3D-VLA: A 3D Vision-Language-Action Generative World Model

Year: 2024 | Citations: 202 | Authors: Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan

Abstract

Recent vision-language-action (VLA) models rely on 2D inputs, lacking integration with the broader realm of the 3D physical world. Furthermore, they perform action prediction by learning a direct mapping from perception to action, neglecting the vast dynamics of the world and the relations between actions and dynamics. In contrast, human beings are endowed with world models that depict imagination about future scenarios to plan actions accordingly. To this end, we propose 3D-VLA by introducing a new family of embodied foundation models that seamlessly link 3D perception, reasoning, and action through a generative world model. Specifically, 3D-VLA is built on top of a 3D-based large language model (LLM), and a set of interaction tokens is introduced to engage with the embodied environment. Furthermore, to inject generation abilities into the model, we train a series of embodied diffusion models and align them into the LLM for predicting the goal images and point clouds. To train our 3D-VLA, we curate a large-scale 3D embodied instruction dataset by extracting vast 3D-related information from existing robotics datasets. Our experiments on held-in datasets demonstrate that 3D-VLA significantly improves the reasoning, multimodal generation, and planning capabilities in embodied environments, showcasing its potential in real-world applications.